

LIMITR: Leveraging Local Information for Medical Image-Text Representation (Supplementary Material)

1. Phrase-grounding

For phrase-grounding evaluation (on MS-CXR dataset), we followed [3] and trained our model using only the impression section of the reports. For this task, we dropped all the images of MS-CXR from the training set, to make sure that our model uses them only for inference. Qualitative results presented below for each of the 8 abnormality classes in MS-CXR dataset (Figure 2).

2. Cross modal alignment-extension

Throughout the paper, for simplicity, we describe how our method creates and utilizes weighted visual representation with respect to each textual feature. We hereby extend the explanation from the paper and describe the complementary direction – weighted textual features with respect to each image region.

Cross modal alignment. This section is extension of section 3.2 from the paper. To create the weighted textual features with respect to each image region we start with computing the cosine similarity c_{ij} between v_i and t_j , to create $c_i = [c_{i1}, c_{i2}, \dots, c_{iN_w}]$. It is further normalized using softmax, in order to get an attention weight:

$$w_i = \text{softmax}(\lambda c_i). \quad (1)$$

The attended visual feature m_i with respect to the i^{th} image region is the weighted sum of all the textual local representations:

$$m_i = \sum_{j=1}^{N_w} w_i[j] \cdot t_j. \quad (2)$$

Next, we calculate the alignment between v_i and its corresponding m_i . The local alignment a_i is calculated as:

$$a_i = \mathcal{A}(m_i, v_i) = \frac{m_i \circ v_i}{\|m_i \circ v_i\|_2}, \quad (3)$$

where \circ is an element-wise multiplication and $\|\cdot\|_2$ is the L_2 -norm.

Aggregation. This section is extension of section 3.3 from the paper. The local alignments are aggregated into a global alignment vector using a weighted sum. Let the local

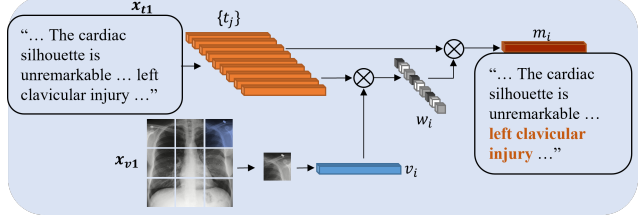


Figure 1: **Cross-modal alignment- textual weighted representation.** Given an image-report pair (x_v, x_t) , we compute for each image region v_i its corresponding textual weighted representation m_i . This is done by using the similarity between each word representation t_j to v_i as the weight for this region $w_i[j]$. The right text shows that the words that correspond to the selected image region are highlighted in orange, representing the higher weight of these words. The final textual representation, m_i , is created by a weighted sum of t_j . This figure corresponds to figure 3 from the paper.

alignments, computed at the alignment module, be $\mathcal{A}_T = \{a_1, a_2, a_3 \dots a_{N_r}\}$. Let \bar{a} be the mean of \mathcal{A}_T . The weight of a_t is defined as:

$$q_t = \left(\text{softmax} \left(\frac{W_q \bar{a} \cdot (W_k \mathcal{A}_T)^T}{\sqrt{d}} \right) \right)_t, \quad (4)$$

where $1 \leq t \leq N_r$, W_q and W_k are linear transformations of the self-attention, and d is the feature dimension.

The final alignment vector between an image and a report is defined as:

$$a_f = \sum_{t=1}^{N_r} q_t (W_v \cdot a_t). \quad (5)$$

Loss. This section is extension of section 3.4 from the paper. Recall that our loss is composed of three components: global, local internal and local external. The use of weighted textual features influence only local alignments, therefore we hereby describe the local internal and local external losses which use the weighted textual features with respect to each image region.

Local external loss: the loss is computed as described in equation (8), while this time \mathcal{A}_{agg} is the aggregated representation based on the weighted-textual representations and the visual representations, as described in the section above.

Local internal loss: L_{int} , is given the local visual representation, v_i , as well as its corresponding attention weighted textual representation, m_i .

The loss is defined as follows:

$$L_{int}(x_v^k, x_t^k) = \sum_{i=1}^{N_r} (l_k^{v_i|m} + l_k^{m_i|v}),$$

$$l_k^{v_i|m} = -\log \left(\frac{\exp(a_i(v_i, m_i)/\tau)}{\sum_{j=1}^{N_r} \exp(a_i(v_i, m_j)/\tau)} \right), \quad (6)$$

$$l_k^{m_i|v} = -\log \left(\frac{\exp(a_i(v_i, m_i)/\tau)}{\sum_{j=1}^{N_r} \exp(a_i(v_j, m_i)/\tau)} \right).$$

We sum the losses related to the weighted visual features with the losses related to the weighted textual features for both the local internal loss and the local external loss.

3. Implementation details

Training data pre-processing (MIMIC-CXR). Images: we resize the original images to 256 pixels on the larger dimension and randomly/center cropped (training/inference) the resized images. We used Resnet50 as our image encoder. We extract the local features from the last convolution layer to get 19X19 feature maps which result in 361 local image regions.

Reports: we extract the impression and findings sections of the report using the official script provided in the github of MIMIC-CXR. We used Bio-clinical BERT tokenizer to tokenize the extracted text. Following [8] we set the maximum number of tokens per report to 97.

Training details. We used Adam optimizer with an initial learning rate of 5e-5 and a weight decay of 1e-6. We measured the performance on the validation set using $Recall@K$ where $K=1,5,10$ for both Image-to-Text retrieval and Text-to-Image. We set the maximum number of epochs to 50 and used the sum of those metrics R_{sum} as the early stopping criteria (plateau over 5 epochs). The best checkpoint is defined as the checkpoint with the highest R_{sum} . We use batch size of 48. All models were trained on a single A100 GPU. For all losses we used $\tau = 0.1$.

4. Additional results

Classification. Following previous works [8,23,27], we employ the Linear Classification framework to assess the transferability of our trained image encoder. This involves the freezing of the trained ResNet-50 image encoder while exclusively training a randomly initialized linear classification head for the subsequent classification task. We as-

sess our model’s performance on the CheXpert and RSNA datasets using 1%, 10%, and 100% of the training data. We report the area under the ROC curve (AUROC) as the evaluation metric for both datasets (Table 1). The results of [8,23,27] are reported in their respective papers. We outperform their results for the described task.

| Method | CheXpert | | | RSNA | | |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 1% | 10% | 100% | 1% | 10% | 100% |
| MGCA [23] | 87.6 | 88.0 | 88.2 | 88.6 | 89.1 | 89.9 |
| ConVIRT [27] | 85.9 | 86.8 | 87.3 | 77.4 | 80.1 | 81.3 |
| GLoRIA [8] | 86.6 | 87.8 | 88.1 | 86.1 | 88.0 | 88.6 |
| Ours | 88.0 | 88.3 | 88.7 | 88.7 | 89.4 | 90.6 |

Table 1: **Classification.** Linear classification results (AUROC [%]) on CheXpert and RSNA with 1%, 10%, 100% training data.

Aggregation weights visualization. As described in section 3.3 in the paper, we assign a weight q_t to each local alignment a_t based on their level of informativeness. Our goal is to assign higher weights to local representations (words from the report) that describe distinct information, such as the pathologies.

Below are three examples of reports along with the corresponding weight assigned to each word in the report.

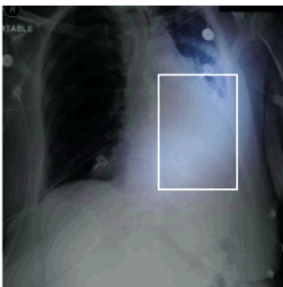
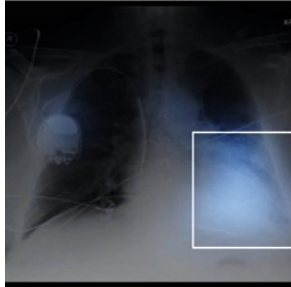
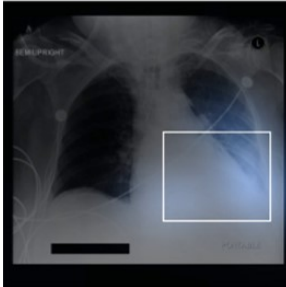
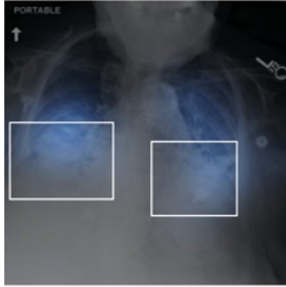

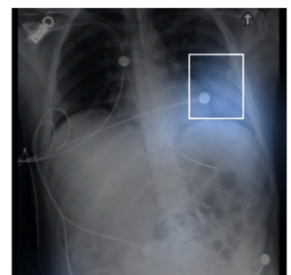
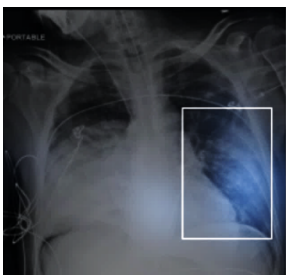

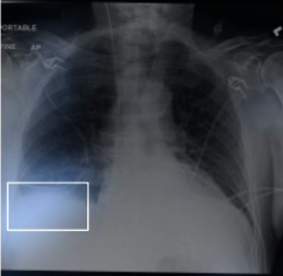
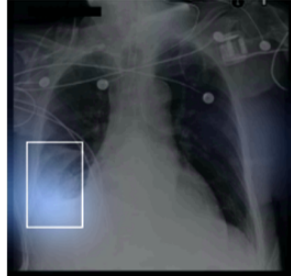
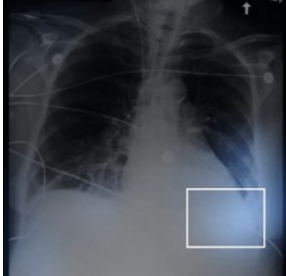
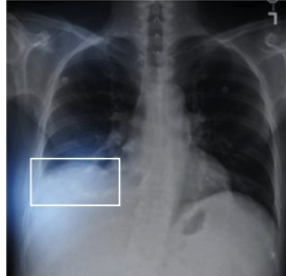

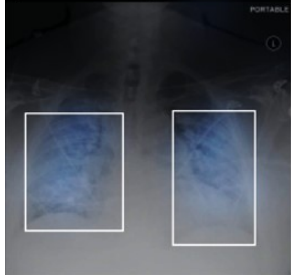


Each word is shaded to reflect its weight (value of q_t); the darker the color the higher the weight. As expected, the mentions of the pathology receive higher weights compared to descriptions of normalities, such as the size of the cardiac silhouette. Each example is marked with a different color.

“... there is moderate tortuosity of the thoracic aorta ... the lungs appear clear ...”

“... there is no pneumothorax ... bilateral pleural effusions are likely ...”

“... grossly unchanged bilateral pleural effusions ... cardiac silhouette is within normal limits ...”

Comparison to pre-trained models. Pre-trained models that were trained on massive amounts of image-text pairs in the natural image domain (i.e., CLIP) do not generalize well to the medical domain. For example, in image-to-text retrieval by CLIP, a generic sentence about the presence of pneumonia in a radiograph is identified as the top-1 match of 60% of the images, despite only 7% of the images actually containing Pneumonia. Only adapting the training data is not enough. [27] presents a similar approach to CLIP with adaptations for the medical domain, by replacing the encoders and the training data. As shown, Our results outperform [27]’s (Tables 1-3 in the paper).

| | | | | | |
|-------------------------|---|---|---|--|---|
| Atelectasis | <p>"Left lower lobe collapse is new"</p> <p>CNR 1.625</p> | <p>"possible small left pleural effusion with adjacent atelectasis"</p> <p>CNR 1.932</p> | <p>"left lower lobe is still collapsed"</p> <p>CNR 1.663</p> | <p>"multisegmental lower lobe opacities are present, consistent with areas of atelectasis lung"</p> <p>CNR 0.909</p> | |
| |  |  |  |  | |
| | Lung Opacity | <p>"hazy opacity in the left suprahilar lung may represent ground-glass opacity"</p> <p>CNR 1.877</p> | <p>"patchy ground-glass opacities are seen in the left lung base"</p> <p>CNR 1.599</p> | <p>"ground-glass opacities in the left lung"</p> <p>CNR 1.413</p> | <p>"patchy ground-glass opacity in the mid right lung is also present"</p> <p>CNR 2.897</p> |
| | |  |  |  |  |
| Pleural effusion | | <p>"small pleural effusion is stable"</p> <p>CNR 2.507</p> | <p>"moderate right pleural effusion is unchanged compared to the prior exam"</p> <p>CNR 1.900</p> | <p>"the minimal left pleural effusion persists"</p> <p>CNR 1.658</p> | <p>"small amount of associated right pleural effusion is demonstrated"</p> <p>CNR 1.602</p> |
| | |  |  |  |  |
| | Edema | <p>"evidence of worsening pulmonary edema and mitral regurgitation"</p> <p>CNR 1.832</p> | <p>"hazy bilateral parenchymal opacities favored to represent edema"</p> <p>CNR 1.378</p> | <p>"hazy perihilar opacities maybe due to pulmonary edema"</p> <p>CNR 1.698</p> | <p>"interstitial edema is present in the right lower lung"</p> <p>CNR 1.810</p> |
| | |  |  |  |  |

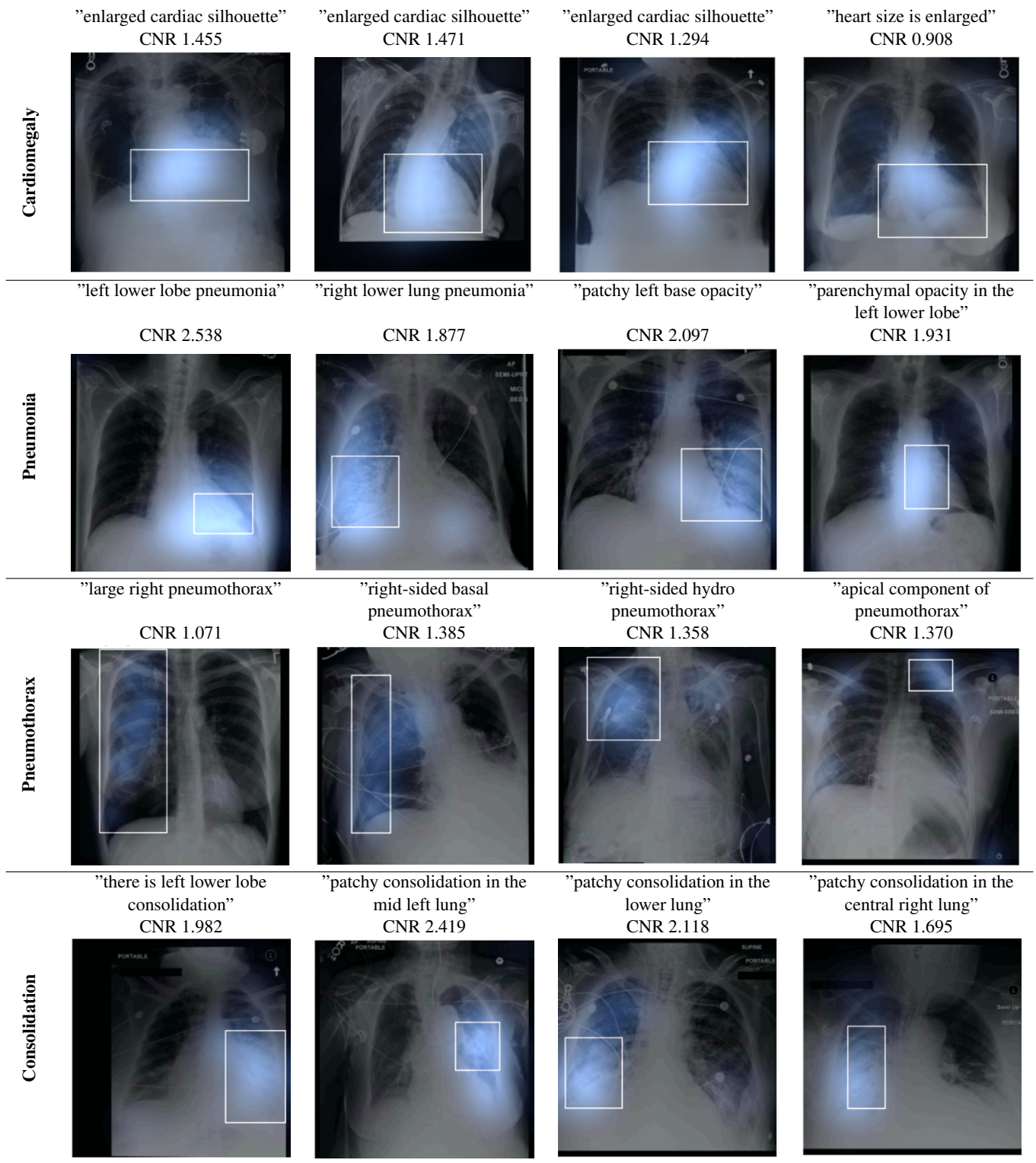


Figure 2: **Phrase-grounding — qualitative results.** Given a phrase and an image, the goal is to produce a similarity map between the phrase and the image. Brighter color indicates higher similarity of that region to the given phrase. The results are measured using CNR; higher values indicate good localization of the phrase in the image. Each row in the figure represents one of the abnormality classes of MS-CXR dataset. Recall that our model was not trained for that task, but still learned to match well between image regions and their corresponding phrases.